Visual perception of unitary elements for layout analysis of unconstrained documents in heterogeneous databases

Baptiste Poirriez, Aurélie Lemaitre, Bertrand Coüasnon

Visual perception of unitary elements for layout analysis of unconstrained documents in heterogeneous databases

Baptiste Poirriez*, Aurélie Lemaitre† and Bertrand Coüasnon*
*Irisa - INSA
† Irisa - Université Rennes 2
Université Européenne de Bretagne, Rennes, France
Email: firstname.lastname@irisa.fr

***Abstract*—The document layout analysis is a complex task in the context of heterogeneous documents. It is still a challenging problem. In this paper, we present our contribution for the layout analysis competition of the international Maurdor Campaign. Our method is based on a grammatical description of the content of elements. It consists in iteratively finding and then removing the most structuring elements of documents. This method is based on notions of perceptive vision: a combination of points of view of the document, and the analysis of salient contents. Our description is generic enough to deal with a very wide range of heterogeneous documents. This method obtained the second place in Run 2 of Maurdor Campaign (on 1000 documents), and the best results in terms of pixel labeling for text blocs and graphic regions.**

***Keywords*-document layout analysis, heterogeneous documents, business documents, tables, forms**

I. Introduction

This work addresses the problem of document layout analysis of heterogeneous and unconstrained documents. In an industrial context, automatic document processing has to face with a large variety of documents: there is no a priori about the possible content of the documents. The actual methods of document layout analysis are often trained to recognize a specific kind of document but the processing of heterogeneous databases of documents is still an open problem. Moreover, the modern layouts of documents can vary and it is not possible to predict a standard document organization.

In this context, an international competition called *Maurdor Campaign*[1] was led in November'2013 (second run). One of the tasks was the layout analysis of heterogeneous documents, such as those presented on figure 1. The images contain mixed printed and handwritten text, with three languages (French, English, Arabic). They sometimes contain graphics, logos, tables, forms... The goal was to localize homogeneous text regions (script and language), tables, and various kinds of graphic regions.

In this paper, we present our system for this competition. It is based on a grammatical description of the possible contents of the documents. The originality is to iteratively look for the most salient and structuring elements in the document, and to authorize a new segmentation of the document during the analysis.

This paper is organized as follows: after a study of the related works, our method will be presented on section III. Then, we will describe how we have implemented our system with the existing DMOS-P method. We will end in section V with the results that we have obtained in Maurdor Campaign and with a discussion on possible improvements.

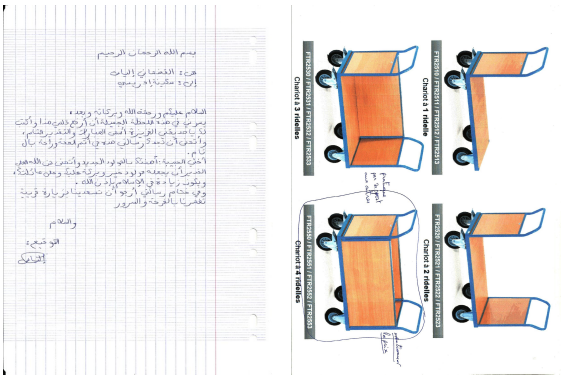
II. Related work

Document understanding presents various challenges [2]. In this work, we have to deal with two difficulties: the heterogeneity of the documents and the heterogeneity of the database. Those two aspects are often studied separately in the literature.

We call heterogeneous documents some documents that may contain various kinds of contents: handwritten text, printed text, graphics, tables, forms. Some works are proposed for pixel labeling of heterogeneous documents inside of a quite homogeneous collection. Recently, Cote *et.al.* [3] propose a method to label pixels in business documents. This approach is based on texture analysis, as many other approaches cited in their state of the art. Those approaches are convenient when we can rely on certain stability inside of the collection. For example, on the proposed business document collection, the text blocs are quite homogeneous.

On the other hand, the study of heterogeneous databases often leads to a problem of classification. For example, Medvet *et. al.* [4] propose to analyze the layout of various kinds of printed documents. However their method requires learning a new class for each new kind of document. Many methods have been proposed for the classification of document image in heterogeneous images, as summarized in this survey [5].

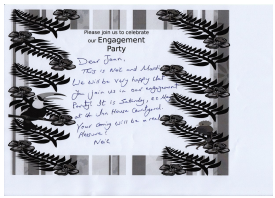
In our case, we do not know a priori the kind of document that can be submitted to the layout analysis process. Then it is not possible to learn classes of documents. We have to focus on the analysis of structuring elements that might be present. In that context, the existing work often focus on one kind of element. For example, in [6], the authors extract tables in documents with varying layout (company



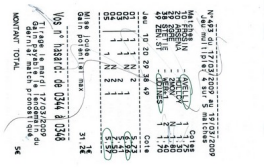
(a) Personal handwritten letter in Arabic (b) Page of catalog, annotated by hand

(c) Flyer

(d) French completed form



(e) Invitation card with graphics



(f) Commercial printed bill

Figure 1. Unconstrained database of heterogeneous documents

report, newspapers, magazines...). An extraction of tables is also proposed by LITIS lab[7] on the Maurdor database.

The particularity of this work is to combine the two difficulties: heterogeneous contents of documents and unconstrained databases. To our knowledge, there are few studies in the literature dealing with those two difficulties in the same time, except for Maurdor database [8].

III. Our method for Document Layout Analysis

In this section, we present our work for layout analysis, which begins after a pre-processing step.

A. Pre-processing: orientation detection

Due to the heterogeneity of the database, we have to apply a preprocessing tool to detect the orientation of the documents. Our strategy is to use an element of perceptive vision: when having a global vision of a document, a human can detect the main orientation of a document (with a precision of 180°). The human vision is guided by the main direction of the writings. We use this idea for the orientation detection: we extract the line segments in a low resolution image (built by sub-sampling), which give indications on the document orientation. Indeed, the text-lines are perceived by the human eye as line segments at low resolution. If the number of vertical line segment is significantly higher than the number of horizontal line segment, then we rotate the document.

B. Overview of the method

We have to detect structuring elements, without knowing if they are present in the document. We propose to use two principles of perceptive vision that we used in the context of homogeneous databases [9]:

- some contents are salient for the human vision inside of a document, which are often strongly structuring;
- combining several points of view of a document enables a prediction/verification mechanism: the layout is predicted in a global vision of a document and verified with details.

Our method is based on a grammatical description of the content. The strategy of analysis consists in iteratively finding the most structuring and salient elements: the tables and boxes, then the text blocs and at last the graphics. The originality of this analysis is to ask for a new segmentation of the document, during the analysis, to take into account the previously detected elements.

The figure 2 presents an overview of the different steps of analysis. Each step is detailed in the following sections.

Figure 2. Overview of the proposed method

We use several kinds of primitives as terminal elements of our grammar:

- the line segments,
- the connected components,
- the printed words that are recognized by an OCR.

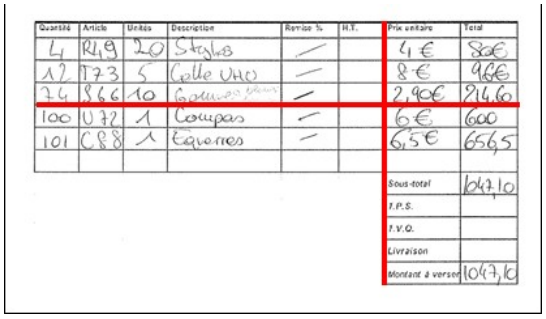
Depending on the level of analysis, those elements are detected at different resolution levels of the image.

C. Tables, boxes, separators

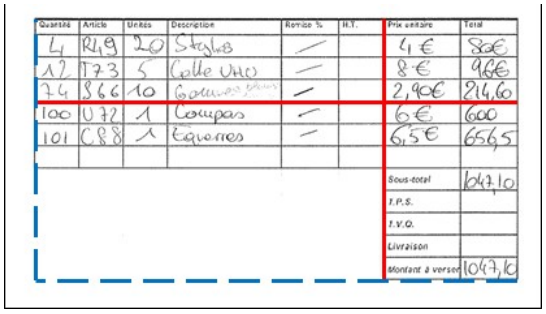
The first part of our work consists in localizing tables, boxes and separators. Those elements are based on the presence of rulings. Consequently, we use the results of a line-segment detector (based on Kalman filtering) as an input of our grammatical description.

We first consider that a document can contain tables. A table must be composed of at least two crossing rulings

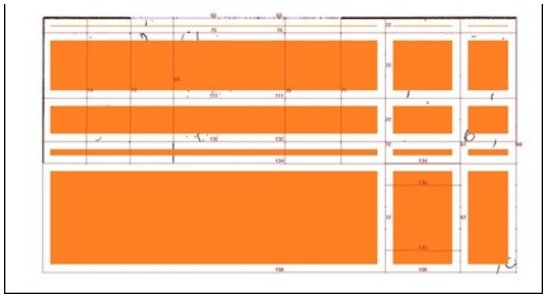
(figure 3(a)). Then we look for parallel horizontal or vertical rulings having the same size. However, sometimes all the rulings are not present, or the tables are not rectangular, and it is not possible to find parallel rulings having the same size as the initial cross. Consequently, we compute some "virtual rulings" (figure 3(b)) that has the same length as the base crossing rulings. Once each ruling is detected, we compute the cells inside of the table (figure 3(c)).



(a) Localization of the base crossing rulings (in red)



(b) Construction of virtual rulings (in dotted blue) to deal with not-rectangular tables



(c) Final cells in which is called a recursive analysis

Figure 3. Table localization

Our grammatical description expresses that we have to recursively analyze the content of the cells of a table or of the boxes. Indeed, inside of a cell, we can find another table or any constituent of a whole document. This method enables to easily deal with recursive tables.

The isolated boxes are made of four isolated rulings that constitute a rectangle. The remaining long enough rulings are considered as separators.

D. Latin printed text analysis

For printed text analysis, we use the commercial OCR Abbyy FineReader CLI. This OCR does not provide a correct layout on very unconstrained heterogeneous documents, but it enables to localize some printed words (figure 4(a)), even if documents also contain handwritten text and graphics.

Those words are used as input of our grammatical description. We select the words that we can trust, depending on various criteria given by the OCR: belonging to a dictionary, high enough recognition confidence, size of the word. These criteria enable to filter the handwritten words that are detected as printed by the OCR. The grammatical rules then describe a text bloc as a set of words that are organized into lines (figure 4(b)). The grammar checks among others the text alignment, the consistence of font-size. It also detects the presence of columns to build the final text-blocks.

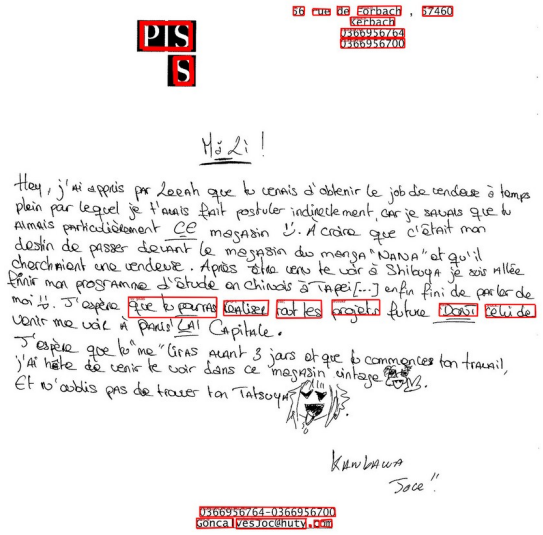
E. New segmentation

At this step, the most structuring elements of the document have been detected: tables, boxes and printed text blocs. However, those elements sometimes interfere with the detection of the remaining elements (handwritten text or graphics). For example, in a form field, the ruling may cross the handwriting, which causes troubles to detect the connected components that compose the handwriting.

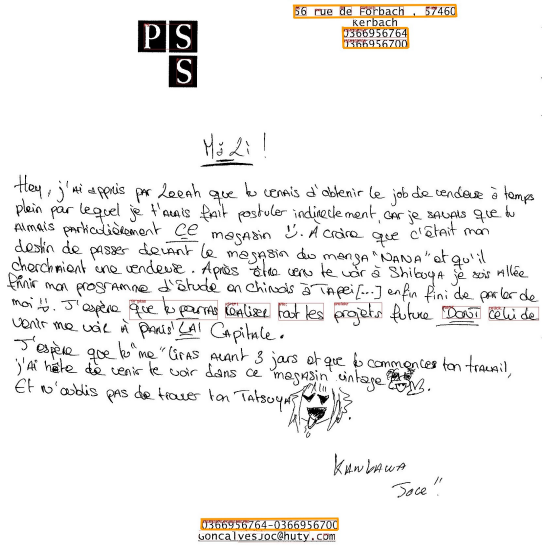
In order to deal with this problem, we propose to remove the previously detected elements before re-segmenting the analysis. The "removing" is done by replacing in the image all the black pixels by a gray level that corresponds to the estimated level of the local background color of the document. Then, the new extraction of connected components is realized.

F. Text-line extraction

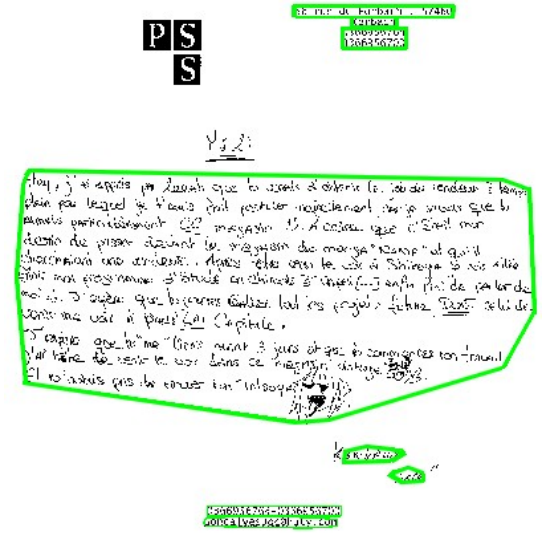
The next step consists in extracting text lines in documents that now contain mixed graphic and text lines. The remaining text lines can be Latin or Arabic handwritten, Arabic printed and even the Latin printed text lines that have not been detected by the OCR. Our description of text lines must be robust enough to deal with all those cases. This is realized thanks to a mechanism of perceptive vision[10]: at low resolution, the text lines can be perceived as line segments. We then apply our line segment detector to compute a prediction on the position of text line. Then, in the full image resolution, we can verify the presence of the text lines with the analysis of connected components: a text line is described as a set of regular aligned connected components. The difficulty is to sort the connected components between graphics and text lines when they overlap. We use thresholds on the size of the connected components to determine if they belong to a text-line or to a graphic region. This method can deal with slightly skewed text.



(a) Words located by the commercial OCR



(b) Construction of text blocs with reliable words of OCR



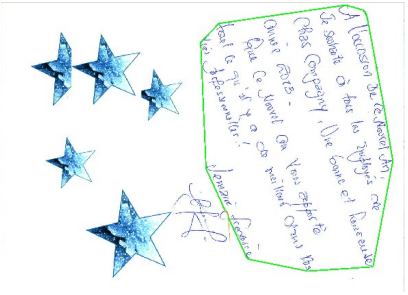
(c) Final text blocs after handwritten text analysis

Figure 4. Text analysis based on OCR

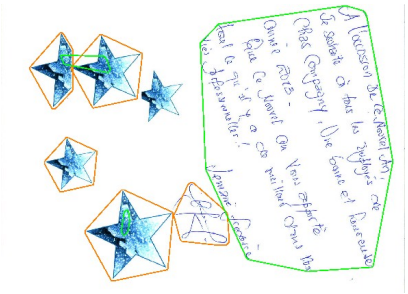
Once the text-line have been localized, they are gathered into text-blocs according to proximity and alignment criteria. Some examples of final text-blocs are presented on figure 4.

G. Graphic localization

The last step of analysis consists in localizing the graphic regions among the remaining components of the image. We use again a concept of the perceptive vision: the graphic regions are perceived as salient elements when we have a global view of the documents. Consequently, we work in a low resolution image (dimensions divided by 4 by sub-sampling). We consider that a graphic element is a big enough (more than 5 pixel edgewise) connected component in this image (figure 5).



(a) Initial document after text-bloc detection



(b) Detection of salient graphics

Figure 5. Graphic detection (a signature is a graphic element)

IV. Implementation of our method

We have implemented our approach using the DMOS-P method.

A. DMOS-P method

DMOS-P (Description MOdification of the Structure with Perceptive vision) [11] is a grammatical method for the recognition of structured documents. It is based on a specific grammatical language, EPF, which enables to express the physical, syntactical and semantic organization of a kind of document. Once the grammatical description has been realized, the associated parser is automatically produced by a compilation step.

The DMOS-P method has been validated on many kinds of documents (music scores, tables, forms, archive documents, mathematical formulas, business letters) and at a

large scale (more than 700,000 document images). However, it has always been applied to relatively homogeneous databases.

B. Our implementation

In this work, the novelty is that we have to use DMOS-P method in the context of heterogeneous documents and unconstrained database. Consequently, we cannot describe precisely the content of each kind of document. Note that we had previously work on certain components of the documents: tables[11], text-lines[10]. But as it was on a homogeneous context, our descriptions were not adapted for heterogeneous documents. We had to adapt them to describe all the unitary elements according to the strategy presented in section III. It was possible thanks to the powerful expressiveness of the EPF language. As the parser can deal with the presence or absence of each of these elements, this grammatical description is common for all the documents, even if they are heterogeneous.

Moreover, the DMOS-P parser is particularly adapted to this problem as it enables to ask for a new segmentation of the document during the analysis, which is required once we have detected the tables and rulings. This new segmentation is led by the grammatical description, which describes when it is necessary to remove elements.

At last, we have to combine several kinds of primitives: connected components, line segments, OCR words, extracted at various resolutions. This is very simple in DMOS-P method that offers the concept of *perceptive layer*. Indeed, each kind of primitive is stored in a layer and the grammatical description expresses which is the layer to study at each step of the analysis. Thus, the contents of the layers can be combined to produce some more complex results.

This implementation has been validated in the context of the international Maurdor Campaign. An example of produced result is presented on figure 6.

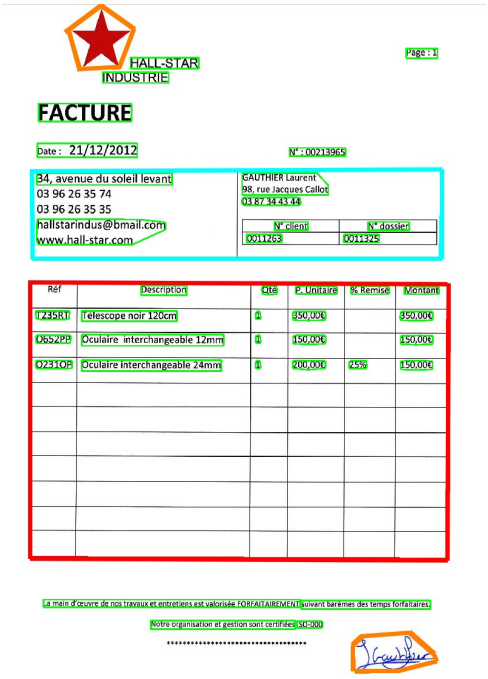


Figure 6. Example of final segmentation: box in cyan, table in red, graphics and signature in orange, text blocs in green. This image obtains a score of 9.1% error rate with ZoneMap metric

V. Validation with Maurdor Campaign

The Maurdor Campaign[1] is an international competition that aims at evaluating the various steps of document processing. It was led in November'2013 by the French lab LNE, Cassidian - an EADS company, and funded by the DGA. Several tasks were proposed in this campaign. We focus on the first one, called "Module 1", which evaluates the document layout analysis step.

A. Database and Metrics

The participants of the campaign were provided two sets of annotated data: a train set of 6,127 documents and a dev set of 1,000 documents. The competition results were evaluated on a third set: a test set of 1,000 documents. The database contains various kinds of documents (figure 1): blank forms and completed forms, typewritten commercial documents, handwritten personal letters, commercial letters, maps, newspaper articles... Three languages are present: French, English and Arabic.

The documents have been manually labeled, following a guide of annotation. The annotations mainly contain homogeneous text blocs (with the same language and the same script), table regions, boxes, separators and various kinds of graphic regions (including signatures).

The evaluated systems must provide the best polygonal zones enclosing the regions of the ground truth, with the correct type. Two metrics have been used in the competition. The ZoneMap metric evaluates the quality of segmentation in terms of split and merge of polygonal zones. It is an error rate (the smaller is better). The Jaccard metric evaluates at pixel level if the correct type (text, graphic element, table...) has been assigned to each pixel. The Jaccard score is a success value (the higher is better). Those two metrics are complimentary. They are fully described on the website of the campaign [1].

B. Obtained results

Three participants took part to this campaign. We present the global results of the campaign in table I. Our method obtains the second position in the global rates of the competition. Indeed, we mainly focused on finding the type of the elements in the image, more than building the ideal blocs. Consequently, our ZoneMap score, which evaluates split and merge, is not very good whereas our Jaccard Score, which evaluates to good classification of pixels, is very close to the

best participant.

Participant	ZoneMap(%)	Jaccard
Participant 1	**48.7**	**0.45**
Our method	*59.2*	*0.44*
Participant 2	73.5	0.28

Table I
RESULTS OF MAURDOR CAMPAIGN; LOWER ZONEMAP IS BETTER, HIGHER JACCARD IS BETTER

As our system focuses on some specific elements, such as text blocs, we present in table II the detailed results, on each category, with Jaccard metric. Those results show that our system obtains the best scores for the localization of text zones and graphic zones, at pixel level.

Participant	Text zone	Graphic zone	Table
Participant 1	0.552	0.394	**0.363**
Our method	***0.553***	***0.402***	*0.307*
Participant 2	0.307	0.176	0.174

Table II
RESULTS BY CLASS WITH JACCARD METRIC (HIGHER IS BETTER)

C. Discussion

The obtain results for this campaign show that the layout analysis of heterogeneous documents is a very difficult task. It is still an open topic, and our system meets a lot of confusing cases. We mainly have difficulties when graphics overlap text blocs. For example, the image 1(c) presents a high confusion between graphics and text blocs.

We also have difficulties to build homogeneous text blocs. As shows our bad score in ZoneMap, our method causes too much merge and split. In order to improve that, we need to detect in detail if a character is printed or handwritten and its language. For example, a handwritten field inside of a printed text must be isolated in a specific text bloc.

For that purpose, our future work will be to introduce some classifiers at connected component level to know if we have to aggregate a connected component to the current bloc. For example, we plan to use classifiers that separate printed Latin text vs Other, or Arabic text vs Other, or the different kinds of graphics. This should enable a cooperation between the grammatical symbolic description of the page and the statistic output of the classifiers.

VI. CONCLUSION

In this paper, we have presented our method for layout analysis of heterogeneous and mixed documents in the context of Maurdor Campaign. Our method is based on the following characteristics : a grammatical description of recognition rules based on the combination of points of view, an iterative analysis of most structuring elements which are salient in the document, the ability to re-segment the document during the analysis.

It is a new application of our DMOS-P method that had never been applied on such heterogeneous databases. We exploit its genericity and its ability to deal with elements that can be absent.

The results of Maurdor campaign shows that the layout analysis of heterogeneous documents is still an open problem. Our method is placed at the second rank of the competition for the global metric, and at the first place for the labeling of pixels of text and graphic regions.

The future work will be to introduce several classifiers (for script and language detection) and enrich the grammatical description by statistic features. This should enable to build homogeneous text blocs, which is required for the following steps of document recognition.

ACKNOWLEDGMENT

This work has been conducted in collaboration with Cassidian, an EADS company to study, develop and implement a prototype for automatic recognition of documents, for the French Ministry of Defence (DGA).

REFERENCES

[1] DGA, Cassidian, and LNE. (2013) Maurdor campaign dataset. [Online]. Available: http://www.maurdor-campaign.org

[2] A. R. Dengel, "Making documents work: Challenges for document understanding," in *ICDAR 03*, 2003, pp. 1026–1036.

[3] M. Cote and A. Branzan Albu, "Texture sparseness for pixel classification of business document images," *IJDAR*, pp. 1–17, 2014.

[4] E. Medvet, A. Bartoli, and G. Davanzo, "A probabilistic approach to printed document understanding," *IJDAR*, vol. 14, no. 4, pp. 335–347, 2011.

[5] N. Chen and D. Blostein, "A survey of document image classification: problem statement, classifier architecture and performance evaluation," *IJDAR*, vol. 10, no. 1, pp. 1–16, 2007.

[6] F. Shafait and R. Smith, "Table detection in heterogeneous documents," in *DAS'10*, 2010, pp. 65–72.

[7] T. Kasar, P. Barlas, S. Adam, C. Chatelain, and T. Paquet, "Learning to detect tables in scanned document images using line information," in *ICDAR'13*, Aug 2013, pp. 1185–1189.

[8] P. Barlas, S. Adam, C. Chatelain, and T. Paquet, "A typed and handwritten text block segmentation system for heterogeneous and complex documents," in *DAS'14*, 2014.

[9] A. Lemaitre, J. Camillerapp, and B. Coüasnon, "Interest of perceptive vision for document structure analysis," in *Human Vision and Electronic Imaging XV*, 2010.

[10] ——, "Use of perceptive vision for rulling recognition in ancient documents," in *Proceedings of GREC'2009*, 2009.

[11] B. Coüasnon, "DMOS, a generic document recognition method: Application to table structure analysis in a general and in a specific way," *IJDAR*, vol. 8(2), pp. 111–122, 2006.